

# Using decision theory to derive optimal cut-off scores of screening instruments: an illustration explicating costs and benefits of mental health screening

NIELS SMITS,<sup>1</sup> FILIP SMIT,<sup>1,2</sup> PIM CUIJPERS,<sup>1,2</sup> RON DE GRAAF<sup>2</sup>

1 Department of Clinical Psychology, Vrije Universiteit, Amsterdam, The Netherlands

2 The Trimbos Institute (Netherlands Institute of Mental Health and Addiction), Utrecht, The Netherlands

---

## Abstract

*This paper shows how decision theory can be used to determine optimal cut-off scores on mental health screeners. The procedure uses (a) the costs and benefits of correct and erroneous decisions, and (b) the rates of correct and erroneous decisions as a function of the cut-off score. Using this information, for each cut-off point expected costs are calculated. The cut-off point with the lowest expected costs is the optimal cut-off score. An illustration is given in which the General Health Questionnaire is employed as a major depression screener. Optimal cut-off points are determined for four different contexts: patients, health service providers, society, and mental health researchers. As in these four situations different costs are encountered, different optimal cut-off points were found. Copyright © 2008 John Wiley & Sons, Ltd.*

**Key words:** decision theory, cut-off scores, diagnostic testing, economic evaluation

---

## Introduction

In both mental health research and clinical settings, self-report questionnaires are frequently used to screen for mental health disorders. A respondent who is screened positive on the basis of a major depression screener is suspected to be suffering from major depression and is commonly examined more closely (e.g. using a diagnostic interview). Respondents screened negative are suspected not to suffer from major depression and do therefore not receive a closer examination. To separate respondents into these two mutually exclusive groups, a cut-off point is commonly set for the screener above which a responder receives a positive classification. Because in practice screeners do never show a perfect relationship with the actual mental health state, classification errors occur. To come to a proper value of the cut-off score, researchers usually choose a cut-off

score which somehow minimizes the occurrence of these errors.

In the domains of psychology and medicine, independently of each other, procedures for setting optimal cut-off scores for tests have been developed (see, for example, Hambleton and Novick (1973) for psychology and Kraemer (1992) for medicine). Such procedures explicitly take the costs and benefits of correct and incorrect classifications into account to come to the cut-off with the lowest costs. Most mental health professionals come from the medical or psychological area. However, in the mental health field these procedures are hardly, if ever, used. Instead, sub-optimal heuristics are employed. For example, a cut-off score is set which generates the highest sum of sensitivity and specificity (defined later) coefficients (e.g. Leentjens et al., 2000; Papassotiropoulos et al., 1999). Alternatively, Receiver

Operator Characteristic (ROC, defined later) curves are drawn and the cut-off score which lies closest to the upper left corner is selected (e.g. Herrmann et al., 1996).

In the past it was argued that the determination of costs and benefits of correct and incorrect health related decisions was not possible (e.g. Neufeld, 1977). However, in the last decade the health care environment has been dominated by a philosophy of cost containment and managed care. Consequently, the efficiency of health care resources has been intensively studied (e.g. Beazoglou et al., 1998). For example, the monetary costs of treating mental disorders such as major depression have been calculated (Berto et al., 2000; Greenberg and Birnbaum, 2005; Cuijpers et al., 2007). On the basis of such cost-effectiveness analysis, optimal treatment decisions can be inferred (see, e.g. Granata et al., 1998). Moreover, the screening of such disorders has been evaluated economically as well (see, e.g. McAlpine et al., 2004; Valenstein et al., 2001). In the latter group of research, the validity of screening instruments (in terms of diagnostic accuracy) has been explicitly linked to the costs of mental disorders. Thus, on the basis of this type of research, costs of screening errors can be inferred, and therefore optimal cut-offs for mental health screeners may be determined.

This paper is an illustration of how optimal cut-off scores may be derived for screening instruments by taking into account the costs and benefits of correct and incorrect decisions. It is intended to gain the confidence of mental health researchers who are unfamiliar with this procedure. First, a description of using decision theory to come to optimal cut-off scores is given. Then, an illustration of how cut-off scores can be derived in diagnostic testing. As the costs of clinical decisions may be different in different contexts, cut-off scores are determined from four perspectives: patients, health service providers, society, and mental health researchers.

### Decision theory and optimal cut-off scores

Let us assume that a screening instrument is used to detect major depression in the general population, and that scores on both a screening instrument and a diagnostic interview, representing the true mental health state, are known. Although such an interview is only flawless in theory, it is often referred to as 'gold standard'. Table 1 represents this situation. In Table 1,  $D$  is

**Table 1.** Decision table for the screening situation.  $U$  represents the utility of the outcome of the cell

		Diagnosis $D$		
		$D-$	$D+$	
Screener $S$	$S \geq c$ (Positive)	False Positives $FP_c$ $U_{FP}$	True Positives $TP_c$ $U_{TP}$	$Q_c$
	$S < c$ (Negative)	True Negatives $TN_c$ $U_{TN}$	False Negatives $FN_c$ $U_{FN}$	$Q'_c = 1 - Q_c$
		$P' = 1 - P$	$P$	1

Prevalence =  $P = TP_c + FN_c$   
 Level at cut-off  $c = Q_c = FP_c + TP_c$   
 Sensitivity at cut-off  $c = SE_c = TP_c/P$   
 Specificity at cut-off  $c = SP_c = TN_c/P'$   
 Positive Predictive Value at cut-off  $c = PPV_c = TP_c/Q_c$   
 Negative Predictive Value at cut-off  $c = NPV_c = TN_c/Q'_c$

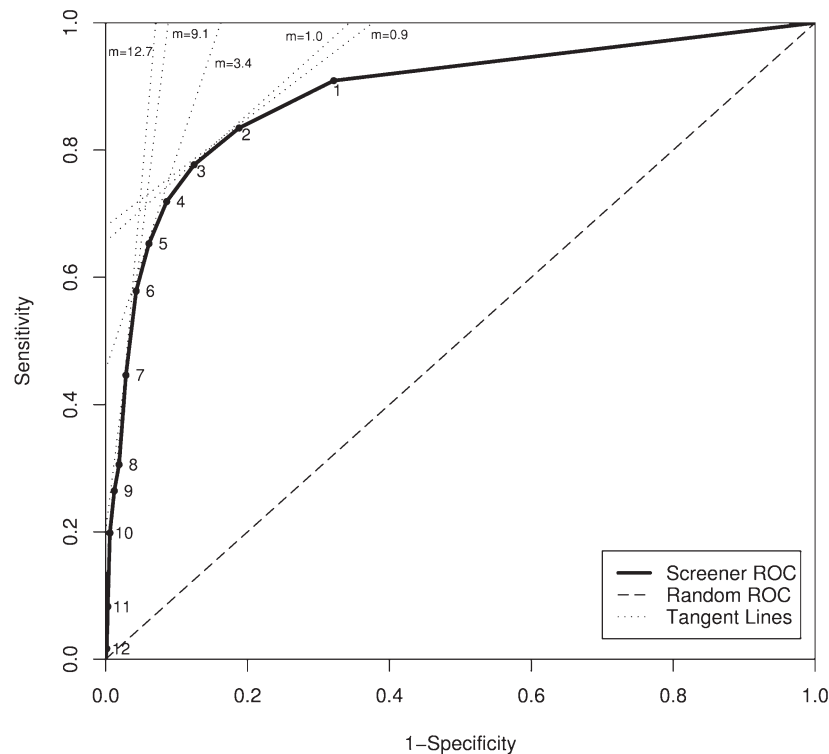
the diagnosis on the basis of the gold standard; a person either does ( $D+$ ) or does not ( $D-$ ) suffer from major depression. The prevalence  $P$  is the proportion of persons actually suffering from major depression. The  $S$  is the score on the screener;  $c$  is a cut-off point on the screener. Persons with scores larger than or equal to  $c$  are screened positive and persons scoring below  $c$  are screened negative. The level  $Q_c$  is the proportion of persons who receive a classification  $S+$ . In practice, due to imperfect validity and measurement errors, screeners never show a perfect relationship with the gold standard. Therefore, as Table 1 shows, there are four possible outcomes: false positives, true negatives, true positives, and false negatives. A false positive error occurs when the screener classifies a respondent to have major depression when, in fact, (s)he has not. A false negative error occurs when the screener classifies a respondent not to have major depression when in fact, (s)he has. The abbreviations ( $FP_c$ ,  $TN_c$ ,  $TP_c$ , and  $FN_c$ ) in Table 1 represent the proportion of the general population in each cell (when using cut-off  $c$ ). It should be noted that these proportions are those in the population sampled, and that  $P$  is a constant for that population. When the population is changed (evaluating Belgians rather than Dutchmen, for example),  $P$ , the four proportions, and the proper cut-off will change as well.

Mostly, the quality of a screener is expressed in terms of two conditional probabilities describing the screener performance with reference to the gold standard. Sensitivity (*SE*) is the probability that a person who suffers from depression is screened as such (see, bottom of Table 1). Specificity (*SP*) is the probability that a person not suffering from depression is screened negative. Alternatively, the quality of the screener is expressed in terms of two conditional probabilities describing the performance of the gold standard with reference to the screener. The Positive Predictive Value (*PPV*) is the probability of a positive diagnosis after a positive screening. The Negative Predictive Value (*NPV*) is the probability of a negative diagnosis after a negative screening.

Commonly, to determine a cut-off score, the *SP* and *SE* of the screening instrument are studied for several cut-off scores. Often, a Receiver Operator Characteristic (ROC) curve (e.g. Zweig and Campbell, 1993) is drawn, placing  $1 - SP$  on the horizontal, and *SE* on the vertical axis. An example of a ROC curve is depicted in Figure 1. *SE* and *SP* are inversely related and vary

with the value of the cut-off score. If a cut-off score is raised, *SE* decreases and *SP* increases ( $1 - SP$  decreases): more respondents not suffering from major depression are correctly screened negative, but fewer depressed patients are detected. In contrast, if the cut-off is set at a lower score, *SE* goes up and *SP* goes down: more respondents who are indeed suffering from major depression are screened positive, but more respondents who are actually not suffering from the disorder are mistakenly screened as positive. When choosing a cut-off with maximal *SE*, the lowest value of *c* should be chosen. When choosing a cut-off with maximal *SP*, the highest *c* should be chosen. Ultimately the choice to maximize *SE* (*SP*) would be to give nearly everyone a positive (negative) classification. Thus, maximizing *SE* or *SP* amounts to not using the screener at all (Kraemer, 1992, p. 68). Apparently to come to a cut-off, the researcher has to take into account the importance of both *SE* and *SP* for the screening situation, but their chosen values should not be too extreme.

A heuristic often encountered in mental health research is to choose that cut-off *c* which has the



**Figure 1.** ROC curve of the GHQ. The numbers on the screener ROC curve represent cut-off scores. The *m* values refer to the slopes of the tangent lines.

maximum sum of  $SE_c$  and  $SP_c$ . It can be shown that this is equivalent to finding the maximum value of the sum  $P' \times TP_c + P \times TN_c$ . Clearly, in this sum the rate of true positives is 'valued' with  $P'$ , and the rate of true negatives is 'valued' with  $P$  (false positive and false negative errors are ignored, and thus have zero importance). If the prevalence is high ( $>0.5$ ),  $TN$  has a higher impact, and if prevalence is low ( $<0.5$ ),  $TP$  has a higher impact in this sum. Only when the prevalence is 0.5,  $TN$  and  $TP$  have equal weight. Clearly, when  $P$  changes (e.g. when screening another population), the importance of a true negative and true positive changes automatically as well. This is undesirable, because it is highly unlikely that a decision-maker's evaluation of classification errors varies from situation to situation. Therefore it is better to explicate the costs and benefits, or utility, of correct and incorrect screening classifications, and incorporate them in a decision theoretical approach.

In decision theory (e.g. Winkler and Hays, 1975, chapter 9), the concept of utility is very important. Utility is a measure of the satisfaction gained from a possible outcome. Utility is often expressed in monetary terms, but may also be expressed on a more subjective scale. It has been introduced to the fields of both psychological and medical testing. In psychological decision-making, Cronbach and Gleser (1965), expressed the validity of selection tests in terms of the dollar value of job performance. Hambleton and Novick (1973) addressed the relationship between cut-off scores and expected utility for mastery tests. Gross and Su (1975) translated this approach to setting the cut-off scores of performance tests in personnel selection. In medical decision making contexts the utility of a test procedure was introduced by McNeil et al. (1975). Kraemer (1988, 1992) and Kraemer et al. (1999) updated these methods. Although in the fields of psychological and medical decision-making, a different terminology is used and seemingly different steps are taken, the procedures for assessing optimal cut-off scores are identical.

To incorporate utility into the diagnostic test setting, the costs and benefits of each of the four outcomes should be specified. This may be expressed in monetary values but also on a more subjective scale. The utility ( $U$ ) associated with each outcome is displayed in Table 1. Subsequently these utilities are weighted by the matching probabilities ( $FP_c$ ,  $TN_c$ ,  $TP_c$ , and  $FN_c$ ) at cut-off  $c$  to calculate *expected* benefits of the four possible outcomes. The expected utility ( $EU_c$ ) for a randomly

selected person is the weighted sum of the four utilities:

$$EU_c = FP_c \cdot U_{FP} + TN_c \cdot U_{TN} + TP_c \cdot U_{TP} + FN_c \cdot U_{FN}, \quad (1)$$

where the probabilities are the weights. By varying cut-off score  $c$ , i.e. by moving the horizontal line in Table 1 vertically, the proportions  $FP_c$ ,  $TN_c$ ,  $TP_c$ , and  $FN_c$  will change. In contrast, the utility of the outcomes remains constant. Consequently, the value of the expected utility  $EU$  in Equation 1 will also change. The optimal cut-off score  $c$  is determined by calculating the expected utility for all possible cut-off scores on the screening instrument, and ascertain which cut-off score has the highest expected utility. Alternatively, when  $U$  is expressed in costs (i.e. disutility) instead of benefits (utility), the optimal cut-off score is the point with the lowest expected costs.

This procedure can be expressed in terms of the ROC method (see, e.g. Kraemer, 1992, p. 121). We start by defining the four classification rates in terms of  $P$ ,  $SE_c$  and  $SP_c$ :  $FP_c = P' \times (1 - SP_c)$ ;  $TN_c = P' \times SP_c$ ;  $TP_c = P \times SE_c$ ; and  $FN_c = P \times (1 - SE_c)$ , where  $P' = 1 - P$ . Then Equation 1 can be expressed as:

$$\begin{aligned} EU_c &= P' \cdot (1 - SP_c) \cdot U_{FP} + P' \cdot SP_c \cdot U_{TN} + P \cdot SE_c \cdot U_{TP} \\ &\quad + P \cdot (1 - SE_c) \cdot U_{FN} \\ &= (P' - P' \cdot SP_c) \cdot U_{FP} + P' \cdot SP_c \cdot U_{TN} + \\ &\quad P \cdot SE_c \cdot U_{TP} + (P - P \cdot SE_c) \cdot U_{FN} \\ &= (P' - U_{FP} + P \cdot U_{FN}) + P' \cdot SP_c \cdot (U_{TN} - U_{FN}) \\ &\quad + P \cdot SE_c \cdot (U_{TP} - U_{FP}). \end{aligned} \quad (2)$$

Since  $P$  and the  $U$ 's do not change when varying  $c$ , the first term above is fixed for all cut-offs. Consequently, only the remaining part of the equation is maximized. Note that  $U_{TN} - U_{FP}$  reflects how much difference in utility it makes whether persons with  $D-$  are classified correctly or not, and that  $U_{TP} - U_{FN}$  reflects how much difference it makes whether persons with  $D+$  are classified correctly or not. Let  $r = (U_{TP} - U_{FN}) / (U_{TN} - U_{FP})$ . Now again, since the  $U$ 's are fixed and thus so is  $r$ , the following is maximized:

$$P' \cdot SP_c \cdot r' + P \cdot SE_c \cdot r, \quad (3)$$

where  $r' = 1 - r$ . This means that maximizing Equation 1 is equivalent to finding that particular cut-off point of the ROC curve, that maximizes Equation 3 for  $P$  and  $r$  fixed. This optimal point occurs where the slope ( $m$ )

of the ROC curve equals  $(P' \times r')/(P \times r)$  (e.g. McNeil et al., 1975, Equation 1). In practice, a line with slope  $m$  that touches the ROC curve is drawn; the cut-off at that touching-point is the optimal cut-off. In Figure 1 several of such tangent lines are drawn.

A problem sometimes encountered is that  $EU$  is maximized when everyone receives a positive ( $Q = 1$ ), or a negative ( $Q = 0$ ) classification, i.e. when the screener actually is not used at all. Therefore, Kraemer (1992) required that  $EU$  at optimal cut-off  $c$  be higher than the random utility ( $RU$ ) for that cut-off.<sup>1</sup>  $RU_c$  can be deduced by introducing a random screener, i.e. a screener which has zero correlation with the gold standard, but with an identical level  $Q$  as the screener. For a random screener,  $FP_c = P' \times Q_c$ ;  $TN_c = P' \times Q'_c$ ;  $TP_c = P \times Q_c$ ; and  $FN_c = P \times Q'_c$ , where  $Q'_c = 1 - Q_c$ .

$$RU_c = P' \cdot Q_c \cdot U_{FP} + P' \cdot Q'_c \cdot U_{TN} + P \cdot Q_c \cdot U_{TP} + P \cdot Q'_c \cdot U_{FN}. \quad (4)$$

Consequently, a valid screener with optimal cut-off  $c$  would require that  $EU_c > RU_c$ .

### Illustration

The illustration employs the 12-item version of the General Health Questionnaire (GHQ; Goldberg, 1972) as a screener of major depression.<sup>2</sup> The range of scores is 0 to 12, with a higher score indicating a higher severity of depression. The diagnostic accuracy of the GHQ was assessed in a large sample ( $N = 5608$ ) from the Dutch population (Bijl et al., 1998), where the CIDI/DSM diagnosis of depressive disorder (one-month recency) was used as the gold standard. Table 2 shows the  $2 \times 2$  classification tables,  $SE$ ,  $SP$ ,  $PPV$  and  $NPV$  for each the 12 cut-off scores. The area under the ROC curve (see, Figure 1), which can be seen as the probability that a randomly selected person with  $D+$  scores higher on the screener than a randomly selected person with  $D-$  (e.g. Zweig and Campbell, 1993), was high:

0.89.<sup>3</sup> When applying the heuristic of finding the highest sum of  $SE$  and  $SP$  to determine the optimal cut-off point, a score of 3 is found. Implicitly, a true positive receives an importance of  $P' = 1 - 0.022 = 0.978$ , and a true negative receives an importance of  $P = 0.022$ ; consequently, a true positive is valued  $0.978/0.022 = 45$  times as important as a true negative.<sup>4</sup>

In the illustration, we will make several assumptions. First, it is assumed that the screening instrument is used to detect major depression in the general Dutch adult population. People either are healthy or have a major depression; other disorders are not addressed. For the determination of costs generated by major depression, we heavily rely on two papers. From the first, a paper by Hakkaart-Van Roijen et al. (2006), we used the estimated average direct and indirect monetary costs generated by patients suffering from major depression. From the second, an article by Valenstein et al. (2001), we used the ratio of costs of undetected to detected major depression. We assume that the outcomes from these two papers are valid for the general Dutch adult population.

In the previous section it was shown that the maximum value of Equation 1 can be derived by looking for the cut-off with slope  $m$  in the ROC curve. We will primarily study Equation 1 rather than the ROC approach for two reasons. First, we think it is more insightful for readers unfamiliar with setting optimal cut-offs to study expected costs than drawing lines in a ROC plot. Second, it allows for displaying  $EU$  relative to  $RU$ , the utility of a random test. However, in the illustration the ROC approach will be touched upon (e.g.  $r$  and  $m$  values are provided). Note again, that both approaches will lead to identical optimal cut-off points.

For each of the four perspectives, the first paragraph starts with an overview of the relevant costs. In the second paragraph, the optimal cut-off score is determined.

### Perspective 1: Respondents

For this perspective we focus on the costs of the screening situation for respondents in the general population.

<sup>1</sup>For a description of the relationship of  $EU$  and  $RU$ , on the one hand, and weighted kappa, on the other hand, see Kraemer (1992) and Kraemer et al. (1999).

<sup>2</sup>The GHQ was originally intended to screen for psychiatric disorders in general. However, for illustrational purposes it is used here to screen for major depression.

<sup>3</sup>The ROC analysis was performed with the ROCR (Sing et al., 2005) library in R (R Development Core Team, 2005).

<sup>4</sup>It can also be shown that using this rule,  $r = P' = 0.978$ , and that  $m = 1$ .

**Table 2.** Diagnostic accuracy of GHQ as a function of cut-off score

Cut-off							Cut-off										
Score	Classification Table			Q	SE	SP	PPV	NPV	Score	Classification Table			Q	SE	SP	PPV	NPV
1	DIAGNOSIS			0.334	0.909	0.679	0.059	0.997	7	DIAGNOSIS			0.038	0.446	0.971	0.256	0.988
	0		1														
	GHQ	1	0.315							0.020							
		0	0.664	0.002													
2	DIAGNOSIS			0.202	0.835	0.812	0.089	0.996	8	DIAGNOSIS			0.025	0.306	0.981	0.262	0.985
	0		1														
	GHQ	1	0.184							0.018							
		0	0.794	0.004													
3	DIAGNOSIS			0.139	0.777	0.875	0.121	0.994	9	DIAGNOSIS			0.018	0.264	0.988	0.323	0.984
	0		1														
	GHQ	1	0.122							0.017							
		0	0.856	0.005													
4	DIAGNOSIS			0.099	0.719	0.914	0.156	0.993	10	DIAGNOSIS			0.010	0.198	0.994	0.421	0.983
	0		1														
	GHQ	1	0.084							0.016							
		0	0.894	0.006													
5	DIAGNOSIS			0.074	0.653	0.939	0.191	0.992	11	DIAGNOSIS			0.005	0.083	0.997	0.370	0.980
	0		1														
	GHQ	1	0.060							0.014							
		0	0.919	0.007													
6	DIAGNOSIS			0.055	0.579	0.957	0.229	0.990	12	DIAGNOSIS			0.002	0.017	0.998	0.182	0.979
	0		1														
	GHQ	1	0.042							0.012							
		0	0.936	0.009													

Note. The Prevalence (*P*) is 0.022.

The costs commonly associated with major depression are decreases in functional status, work capacity, and happiness. To quantify such subjective costs, quality of life research has studied the impact of health states by using preference-based measures (e.g. Wells and Sherbourne, 1999). Such utility ratings are comprehensive measures that take into account all factors influencing the quality of life and arrives at a measurement of patient desirability for a specific health state such as major depression, which is expressed on a scale from 0 (worst) to 1 (best). Following Valenstein et al. (2001) we assume that respondents with *D+* can either show full remission, partial remission or no improvement. In addition, a major depression has a utility of 0.63, a

partial remission has a utility of 0.70, and a full remission has a utility of 0.89. We assume that both depressed persons with and without treatment can show each of these states, but that treated patients have a higher probability of recovery. The proportions in the two groups are taken from Valenstein et al. (2001). Treated patients have a probability of 0.50 of full remission, a probability of 0.30 of partial remission, and probability of 0.20 of no improvement. In contrast, untreated patients have the following probabilities: 0.35 of full remission, 0.15 of partial remission, and 0.50 of no improvement. To come to fair estimates of the average utilities for the untreated patients (false negatives) and treated patients (true positives), the utilities of the four

outcomes are weighted by their relative frequencies. The expected utility of a treated patient with major depression was  $0.50 \times 0.89 + 0.30 \times 0.70 + 0.20 \times 0.63 = 0.78$ . The expected utility of an untreated patient with a major depression was  $0.35 \times 0.89 + 0.15 \times 0.70 + 0.50 \times 0.63 = 0.73$ . The utility of a healthy respondent who is screened negative has the maximum utility of 1.0. The utility of a false positive was set by ourselves. It was argued that both being classified as depressive, and being submitted to a diagnostic interview may decrease the quality of life somewhat. Therefore, this was assumed to have a disutility of 0.01, i.e. a utility of 0.99. Table 3 shows the respondent costs associated with the four outcomes.

Figure 2a shows *EU* and *RU* of the respondents' as a function of GHQ cut-off score. Figure 2a shows that a cut-off score of six generates the highest expected

utility. As the *EU*-curve lies above the *RU*-curve for this cut-off (it lies above it for all values of *c*, although for a cut-off of 12 this difference is marginal), we can conclude that the screener is useful for this situation. In addition, when using the ROC curve to find the optimal cut-off, a slope of  $m = 9.1$  (see Table 3) is used. Figure 1 shows that this line touches the ROC curve at the cut-off of six, the optimal point. In other words, to maximize the average utility of the respondents, a cut-off score of six should be used when administering this major depression screener.

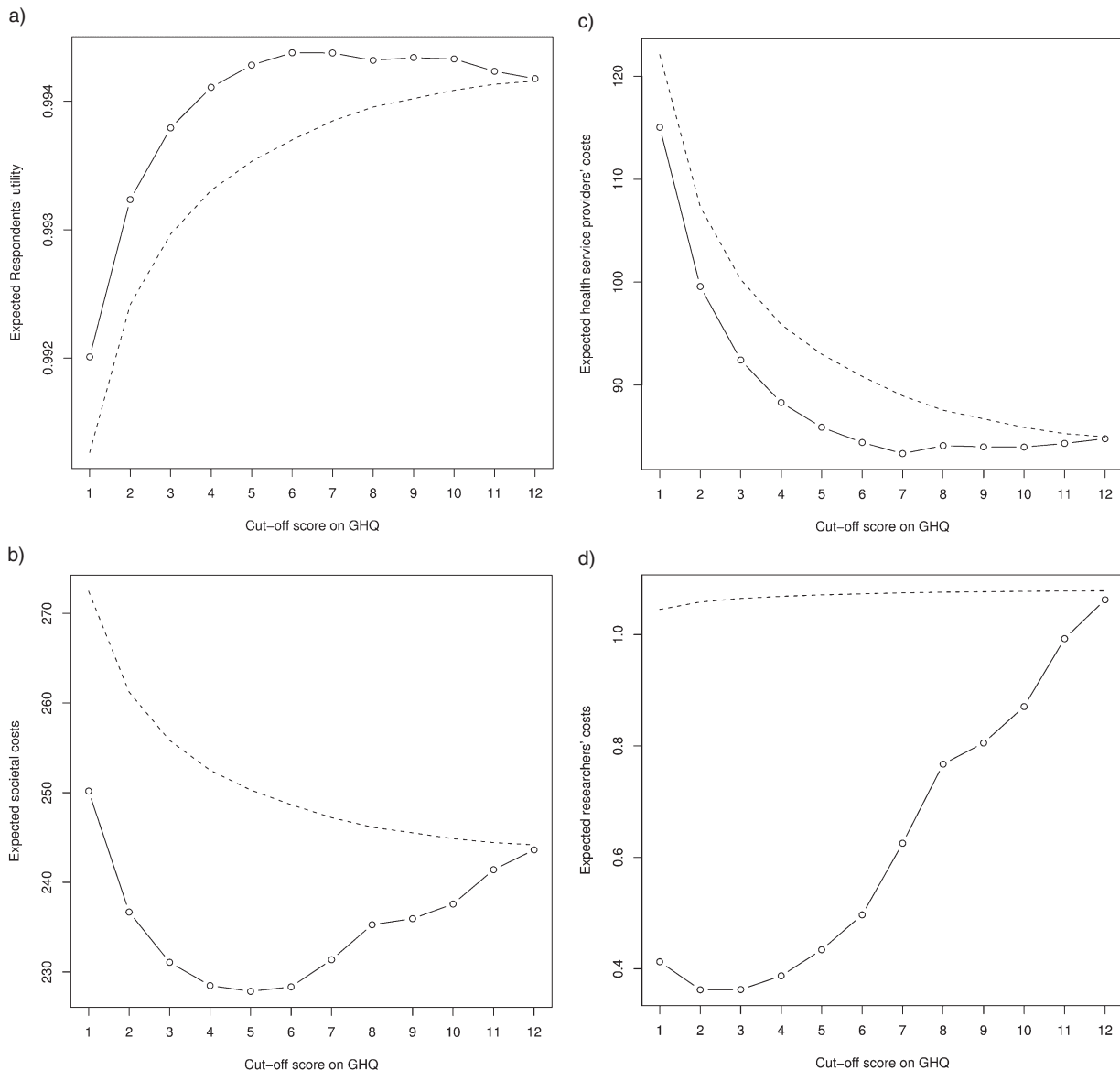
#### Perspective 2: Health service providers

For this perspective we focus on the costs that health care organizations have when dealing with major depression. Several types of costs are made. In the first place, health service providers are faced with the costs of treating major depression. The average per-patient cost of treatment is €1396 (Hakkaart-Van Roijen et al., 2006). In addition, persons suffering from major depression make medical costs other than treating depression as well: €1964 on average. In addition, respondents screened positive are submitted to a diagnostic interview, which costs €124. The total costs of a *true positive* were thus estimated to be  $1396 + 1964 + 124 = €3484$ . The costs of an undetected patient with major depression was determined in the following way. Untreated depressive disorders are associated with high costs for medical care other than treating depression (Hakkaart-Van Roijen et al., 2006; Valenstein et al., 2001). Congruous with Valenstein et al. (2001) we assumed that an undetected, and therefore untreated major depression generates two times as much of such costs as a detected depression. Thus, the cost of a *false negative* is estimated to be  $2 \times 1964 = €3928$ . The cost of a person not suffering from major depression, who is screened as such (i.e. a *true negative*) is assumed to be zero.<sup>5</sup> *False positives* are assumed to only generate the costs of a diagnostic interview (€124). Table 3 shows the health service provider's costs associated with the four outcomes.

**Table 3.** Costs associated with the four perspectives

Costs per cell			<i>r</i>	<i>m</i>																
(a) Respondents' utility																				
<table><tr><td colspan="3">DIAGNOSIS</td></tr><tr><td></td><td>0</td><td>1</td></tr><tr><td rowspan="2">GHQ</td><td>1</td><td><table><tr><td>0.98</td><td>0.78</td></tr><tr><td>1.00</td><td>0.73</td></tr></table></td></tr><tr><td>0</td><td></td><td></td></tr></table>			DIAGNOSIS				0	1	GHQ	1	<table><tr><td>0.98</td><td>0.78</td></tr><tr><td>1.00</td><td>0.73</td></tr></table>	0.98	0.78	1.00	0.73	0			0.833	9.1
DIAGNOSIS																				
	0	1																		
GHQ	1	<table><tr><td>0.98</td><td>0.78</td></tr><tr><td>1.00</td><td>0.73</td></tr></table>	0.98	0.78	1.00	0.73														
	0.98	0.78																		
1.00	0.73																			
0																				
(b) Health service providers' costs																				
<table><tr><td colspan="3">DIAGNOSIS</td></tr><tr><td></td><td>0</td><td>1</td></tr><tr><td rowspan="2">GHQ</td><td>1</td><td><table><tr><td>124</td><td>3484</td></tr><tr><td>0</td><td>3928</td></tr></table></td></tr><tr><td>0</td><td></td><td></td></tr></table>			DIAGNOSIS				0	1	GHQ	1	<table><tr><td>124</td><td>3484</td></tr><tr><td>0</td><td>3928</td></tr></table>	124	3484	0	3928	0			0.782	12.7
DIAGNOSIS																				
	0	1																		
GHQ	1	<table><tr><td>124</td><td>3484</td></tr><tr><td>0</td><td>3928</td></tr></table>	124	3484	0	3928														
	124	3484																		
0	3928																			
0																				
(c) Societal costs																				
<table><tr><td colspan="3">DIAGNOSIS</td></tr><tr><td></td><td>0</td><td>1</td></tr><tr><td rowspan="2">GHQ</td><td>1</td><td><table><tr><td>124</td><td>9635</td></tr><tr><td>0</td><td>11309</td></tr></table></td></tr><tr><td>0</td><td></td><td></td></tr></table>			DIAGNOSIS				0	1	GHQ	1	<table><tr><td>124</td><td>9635</td></tr><tr><td>0</td><td>11309</td></tr></table>	124	9635	0	11309	0			0.931	3.4
DIAGNOSIS																				
	0	1																		
GHQ	1	<table><tr><td>124</td><td>9635</td></tr><tr><td>0</td><td>11309</td></tr></table>	124	9635	0	11309														
	124	9635																		
0	11309																			
0																				
(d) Researchers' costs																				
<table><tr><td colspan="3">DIAGNOSIS</td></tr><tr><td></td><td>0</td><td>1</td></tr><tr><td rowspan="2">GHQ</td><td>1</td><td><table><tr><td>1</td><td>0</td></tr><tr><td>0</td><td>50</td></tr></table></td></tr><tr><td>0</td><td></td><td></td></tr></table>			DIAGNOSIS				0	1	GHQ	1	<table><tr><td>1</td><td>0</td></tr><tr><td>0</td><td>50</td></tr></table>	1	0	0	50	0			0.980	0.9
DIAGNOSIS																				
	0	1																		
GHQ	1	<table><tr><td>1</td><td>0</td></tr><tr><td>0</td><td>50</td></tr></table>	1	0	0	50														
	1	0																		
0	50																			
0																				

<sup>5</sup>In addition, as the screener is administered to subjects in all four cells, and therefore it does not generate differences between cells, the cost of screening itself is not entered in the cost-benefit analyses.



**Figure 2.** Expected (a) respondents' utility, (b) health service provider costs, (c) societal costs, and (d) researchers' costs as a function of GHQ cut-off score. The solid lines represent expected utility (or costs); the dashed lines represent random utility (or costs).

Figure 2c shows the expected health care costs as a function of the GHQ cut-off score. Figure 2c shows that a cut-off score of seven generates the lowest expected costs. In addition, it shows that this optimal cut-off is legitimate since the expected costs are lower than the costs of a random screener. When using the ROC curve to find the optimal cut-off, a slope of  $m = 12.7$  (see Table 3) should be used. Figure 1 shows that this line touches the ROC curve at the optimal cut-off point of seven.

In other words, for the health service provider to keep the costs associated with major depression as low as possible, a cut-off score of seven should be used when administering the screener.

#### *Perspective 3: Society*

From a societal perspective, major depression not only generates direct medical costs. Indirect costs arise when production losses occur due to absence from work

through the major depression. The total indirect costs of a detected major depression is €6151 on average (Hakkaart-Van Roijen et al., 2006). To come to the total societal costs of a *true positive*, the health service provider's costs and indirect costs are added up:  $3484 + 6151 = €9635$ . The indirect costs of an undetected depressive patient are determined as follows. We assume that the degree of recovery from depression is different for patients receiving treatment than for patients receiving no treatment. Following Valenstein et al. (2001) we translated this difference into 20% more indirect costs for a false negative patient. The total average societal costs for *false negatives* are thus  $3928 + 1.2 \times 6151 = €11,309$ . As in the health service provider's scenario, *false positives* only generate the costs of a diagnostic interview (€124). Likewise, the costs of *true negatives* are assumed to be €0. Table 3 shows the societal costs associated with the four outcomes.

Figure 2b shows the expected societal costs as a function of the GHQ cut-off score. The lowest expected costs are encountered for the cut-off score of five. This is a valid optimal cut-off since the expected costs are lower than the costs of a random screener. When using the ROC curve to find the optimal cut-off point, a slope of  $m = 3.4$  (see Table 3) should be used. Figure 1 shows that this line touches the ROC curve at the optimal cut-off point of five. In other words, from a societal perspective, to keep the costs of dealing with major depression as low as possible, a cut-off score of five should be used when administering the GHQ as a screener.

#### *Perspective 4: Researchers*

The fourth perspective is that of mental health researchers. Such researchers often use screening instruments to select subjects from their sample for further research. For example, to select patients for clinical trials, many potential participants are sampled and screened. Screen positives receive a diagnostic interview to determine whether they are suitable to be included in the study (suitable in the sense of having the disorder in question). Obviously, researchers would want the screener to make as little classification errors as possible. However, it is to be expected that they do not assign the same value to false positive as to false negative errors. As the prevalence of mental disorders such as major depression is low, it is rather difficult to obtain enough patients for a study, and therefore undetected patients are quite costly. In contrast, a mentally healthy

person who is screened positive is, although undesirable, not so costly. Such a false positive will soon appear to be unsuitable for the study when administering the diagnostic interview. To quantify this difference, let us assume that for a researcher the costs of a false negative are perceived as 50 times as high as the costs of a false positive. In addition, correct classifications are assumed to generate costs nor benefits. Consequently, as shown in Table 3, the costs of *true negatives*, *false positives*, *false negatives*, and *true positives* are 0, 1, 50, and 0, respectively.

Figure 2d shows the expected researcher's costs as a function of the GHQ cut-off score. The lowest expected costs are encountered for a cut-off score of two (which are marginally lower than for a cut-off score of three). This optimal cut-off is legitimate as the screener gives better results than a random screener. When using the ROC curve to find the optimal cut-off, a slope of  $m = 0.9$  (see Table 3) should be used. Figure 1 shows that this line touches the ROC curve at the optimal cut-off point of two. In other words, for a researcher with the mentioned perceived costs, to keep the costs as low as possible, a cut-off score of two should be used when administering the GHQ as a major depression screener.

#### *Comparing Perspectives*

The four plots in Figure 2 show that the optimal cut-off scores were different for the four perspectives. The optimal cut-off points were six for the respondents, seven for the health service providers, five for society, and two for the researchers. Moreover, the heuristic of choosing the cut-off with the highest sum of *SE* and *SP* had a different optimum as well: a cut-off score of three.

This divergence among the four perspectives resulted from the different ratios of costs encountered (and therefore differences in  $r$  and  $m$ ). For example, a false positive error generated the same costs in the health service provider and societal scenario (€124). However, the ratio of these costs to those of the detected and undetected depressed patients was much higher in the health care scenario than in the society scenario. In the health care situation one merely had to deal with medical costs, whereas in the societal scenario one had to deal with costs due to absence from work as well. Therefore a false negative error added less to the total expected costs in the latter than in the former scenario.

## Discussion

In this paper we showed how optimal cut-off scores on mental health screeners may be derived by incorporating the costs of the potential outcomes. Through weighting these costs by the probabilities of the outcomes, expected costs are calculated. The optimal cut-off point is the score which produces the lowest expected costs (under the condition that the expected costs are lower than those on the basis of a random screener). The illustration employed a major depression screener and used different perspectives for determining costs and benefits: respondents, health service providers, society and researchers. As in these four situations different costs were encountered, naturally, different cut-off points were found to be optimal.

Some may find it disturbing that anything other than the perspective of the patient in evaluating a screening instrument be considered. In practice however, financial resources in mental health are limited and therefore it is impossible to do what is best for all potential mental health patients. Given these restrictions, choices have to be made which are the least costly for mental health professionals. In that context, using optimal cut-off scores for screeners may be helpful. In addition, in different fields of mental health, different costs and benefits are encountered. Therefore, the perspectives of health care providers, society, and researchers were included in the illustration as well.

In the present paper, the costs of all respondents that were in the same cell of the decision table were identical. In decision theoretical words, a 'threshold loss' function was used (e.g. Hambleton and Novick, 1973). Hence, the costs of mild depression were equal to the costs of very severe depression. However, it may be unrealistic to assume that the costs are constant for patients with different degrees of depression. Alternatively, if one wishes to incorporate such differential costs, a linear loss function (Van der Linden and Mellenbergh, 1977) may be used. When using this type of loss function, the costs of misclassifications are linearly related with the degree of depression. Consequently, an undetected very severe depression will generate more costs than an undetected mild depression. Naturally for the application of this function two things should be known. First, the relationship between the screening instrument and the true degree of depression. Second, the costs encountered for the different levels of depression.

In each of the four perspectives we were able to provide costs for each of the four outcomes. Sometimes, it may be difficult to determine these costs precisely. In such cases, attention may be restricted to  $r$ , which quantifies how much difference in utility it makes whether persons not depressed are classified correctly or not, relative to how much difference it makes whether depressed persons with are classified correctly or not (also, see, Kraemer et al., 1999). Although such an estimate may be rather crude, it is still better than not explicating costs and benefits at all.

We showed how optimal cut-offs can be determined on the basis of decision theory. This procedure for setting optimal cut-off points has been available for several decades now. Hopefully, this paper stimulates researchers and clinicians to no longer exclusively use the sub-optimal heuristics for cut-off setting, but to explicate the costs of their clinical decisions in order to set cut-offs that are optimal in the context for which they are used. Moreover, we hope that it makes mental health professionals more aware of the value of the outcomes of screening and improve their use of screeners.

## Acknowledgements

The authors would like to thank the two anonymous reviewers for their very valuable comments.

## References

- Beazoglou T, Heffley D, Kyriopoulos J, Vintzileos A, Benn P. Economic evaluation of prenatal screening for down syndrome in the U.S.A. *Prenatal Diag* 1998; 18: 1241–52.
- Berto P, D'Ilario D, Ruffo P, Di Virgilio R, Rizzo F. Depression: cost-of-illness studies in the international literature, a review. *J Ment Health Policy* 2000; 3: 3–10.
- Bijl R, van Zessen G, Ravelli A. Prevalence of psychiatric disorder in the general population: results of the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Soc Psych and Psych Epid* 1998; 33: 587–95.
- Cronbach LJ, Gleser GC. *Psychological Tests and Personnel Decisions*. Urbana, IL: University of Illinois Press, 1965.
- Cuijpers P, Smit F, Oostenbrink J, de Graaf R, Beekman A. Economic costs of minor depression: a population-based study. *Acta Psychiatr Scand* 2007; 115: 229–36.
- Goldberg DP. *The Detection of Psychiatric Illness by Questionnaire*. London: Oxford University Press, 1972.
- Granata AV, Hillman AL. Competing practice guidelines: using cost-effectiveness analysis to make optimal decisions. *Ann Intern Med* 1998; 128: 56–63.

- Greenberg PE, Birnbaum HG. The economic burden of depression in the US: societal and patient perspectives. *Expert Opin Pharmacol* 2005; 6: 369–76.
- Gross AL, Su WH. Defining a 'fair' or 'unbiased' selection model: a question of utilities. *J Appl Psychol* 1975; 60: 345–51.
- Hakkaart-Van Roijen L, Van Straten A, Al M, Rutten F, Donker M. Cost-utility of brief psychological treatment for depression and anxiety. *Brit J Psychiatry* 2006; 188: 323–29.
- Hambleton RK, Novick MR. Toward an integration of theory and method for criterion-referenced tests. *J Educ Meas* 1973; 10: 159–70.
- Herrmann N, Mittmann N, Silver IL, Shulman KI, Busto UA, Shear NH, et al. A validation study of the Geriatric Depression Scale short form. *Int J Geriatr Psych* 1996; 11: 451–60.
- Kraemer HC. Assessment of  $2 \times 2$  associations: generalization of signal-detection methodology. *Am Stat* 1988; 42: 37–49.
- Kraemer HC. *Evaluating Medical Tests: Objective and Quantitative Guidelines*. Newbury Park, CA: Sage Publications, 1992.
- Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen P, Kupfer DJ. Measuring the potency of risk factors for clinical or policy significance. *Psychol Methods* 1999; 4: 257–71.
- Leentjens AFG, Verhey FRJ, Luijckx GJ, Troost J. The validity of the Beck Depression Inventory as a screening and diagnostic instrument for depression in patients with Parkinsons disease. *Movement Disord* 2000; 15: 1221–24.
- McAlpine DD, Wilson AR. Screening for depression in primary care: what do we still need to know? *Depress Anxiety* 2004; 19: 137–45.
- McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. *New Engl J Med* 1975; 293: 211–15.
- Neufeld RWJ. *Clinical Quantitative Methods*. New York: Grune & Stratton, 1977.
- Papassotiropoulos A, Heun R, Maier W. The impact of dementia on the detection of depression in elderly subjects from the general population. *Psychol Med* 1999; 29: 113–20.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*, 2005.
- Sing T, Sander O, Beerenwinkel N, Lengauer T. *ROCR: Visualizing the Performance of Scoring Classifiers*, Library of the R package, Vienna, Austria, 2005.
- Valenstein M, Vijan S, Zeber JE, Boehm K, Buttar A. The cost-utility of screening for depression in primary care. *Ann Intern Med* 2001; 134: 345–60.
- Van der Linden WJ, Mellenbergh GJ. Optimal cutting scores using a linear loss function. *Appl Psych Meas* 1977; 1: 593–99.
- Wells KB, Sherbourne CD. Functioning and utility for current health of patients with depression or chronic medical conditions in managed, primary care practices. *Arch Gen Psychiat* 1999; 56: 897–904.
- Winkler RL, Hays WL. *Statistics: Probability, Inference and Decision*, 2nd edition. New York: Holt, Rinehart and Winston, 1975.
- Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; 39: 561–77.

*Correspondence:* Niels Smits, Department of Clinical Psychology, Faculty of Psychology and Education, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.

*Email:* n.smits@psy.vu.nl